PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification: C12Q 1/00, C07H 21/02, C12P 19/24, C12P 21/06, C12Q 1/68, G01N 33/48, G01N 33/50

A1

(11) International Publication Number: WO 00/63421

(43) International Publication Date: 26 October 2000 (26.10.2000)

(21) International Application Number: PCT/US00/10484

(22) International Filing Date: 19 April 2000 (19.04.2000)

(30) Priority Data:
60/129,915 19 April 1999 (19.04.1999) US

(60) Parent Application or Grant
BIOS GROUP LP [/]; ( ). KAUFFMAN, Stuart, A. [/]; ( ). LEVITAN, Bennett, S. [/]; ( ). KAUFFMAN, Stuart, A. [/]; ( ). LEVITAN, Bennett, S. [/]; ( ). MORRIS, Francis, E. ; ( ).

Published

(54) Title: A SYSTEM AND METHOD FOR MOLECULE SELECTION USING EXTENDED TARGET SHAPE

(54) Titre: SYSTEME ET PROCEDE DE SELECTION DE MOLECULES UTILISANT UNE FORME CIBLE ETENDUE

(57) Abstract

The present invention is directed to production of a molecule having a predetermined property. In accordance with one embodiment, a library of initial candidate molecules that are at least somewhat dissimilar to a chosen target molecule or "targetshape" is generated. Variants of the initial candidates are generated and screened to identify intermediate candidates from among those variants that are either more or less similar to the targetshape. Data mining techniques such as neural networks are used to extract information about the molecule structures and shapes which lead to the desired activity. Molecular data bases may be screened for candidates matching the preferred shape description or molecules matching the preferred shape description may be synthesized. The process may be iterated by generating variants of the intermediate candidates and screening these variants to identify molecules futher more or less similar to the targetshape.

(57) Abrégé

La présente invention porte sur la production d'une molécule ayant une propriété prédéterminée. Selon une réalisation, on obtient une bibliothèque de molécules candidates de départ qui sont quelque peu dissemblables à une molécule cible choisie ou _ forme cible _. On génère et on recherche des variantes de candidats de départ pour identifier des candidats intermédiaires parmi ces variantes qui sont plus ou moins identiques à la forme cible. Des techniques d'exploration en profondeur de données tels que des réseaux neuronaux sont utilisées pour extraire des informations sur les structures et les formes moléculaires qui mènent à l'activité désirée. Les bases de données moléculaires peuvent être triées pour des candidats correspondant à la description de la forme préférée, ou des molécules correspondant à la description de la forme préférée peuvent être synthétisées. Le procédé peut être répété par génération de variantes des candidats intermédiaires et par tri de ces variantes pour identifier des molécules plus ou moins identiques à la forme cible.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
C12Q 1/00, 1/68, C12P 19/24, 21/06, C07H 21/02, G01N 33/48, 33/50

A1

(11) International Publication Number: WO 00/63421

(43) International Publication Date: 26 October 2000 (26.10.00)

(21) International Application Number: PCT/US00/10484

(22) International Filing Date: 19 April 2000 (19.04.00)

(30) Priority Data:
60/129,915 19 April 1999 (19.04.99) US

(71) Applicant (*for all designated States except US*): BIOS GROUP LP [US/US]; 317 Paseo de Peralta, Santa Fe, NM 87501 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): KAUFFMAN, Stuart, A. [US/US]; 1811 S. Camino Cruz Blanco, Santa Fe, NM 87505 (US). LEVITAN, Bennett, S. [US/US]; 12 Agua Sarca Road, Placitas, NM 87043 (US).

(74) Agents: MORRIS, Francis, E. et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: A SYSTEM AND METHOD FOR MOLECULE SELECTION USING EXTENDED TARGET SHAPE

(57) Abstract

The present invention is directed to production of a molecule having a predetermined property. In accordance with one embodiment, a library of initial candidate molecules that are at least somewhat dissimilar to a chosen target molecule or "targetshape" is generated. Variants of the initial candidates are generated and screened to identify intermediate candidates from among those variants that are either more or less similar to the targetshape. Data mining techniques such as neural networks are used to extract information about the molecule structures and shapes which lead to the desired activity. Molecular data bases may be screened for candidates matching the preferred shape description or molecules matching the preferred shape description may be synthesized. The process may be iterated by generating variants of the intermediate candidates and screening these variants to identify molecules futher more or less similar to the targetshape.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Description

A SYSTEM AND METHOD FOR MOLECULE SELECTION USING EXTENDED TARGET SHAPE

This application claims the benefit under 35 U.S.C. § 119(e) of provisional application number 60/129,915, filed April 19, 1999, which is hereby incorporated by reference in its entirety.

1. INTRODUCTION

The present invention includes a means to obtain one or more initial candidate molecules that are at least somewhat dissimilar to a chosen target molecule or "targetshape", to produce initial variants of the initial candidate molecules, and to screen or select candidates from among those variants for molecules that are either more or less similar to the targetshape.

2. BACKGROUND OF THE INVENTION

The new field of applied molecular evolution, or molecular diversity, is rapidly becoming of central importance in the generation of useful molecules for drugs, vaccines, biosensors, catalysts, and so forth. Molecular diversity is based on generating very large libraries of candidate compounds, up to 10^{15} for quasirandom single stranded RNA or DNA sequences, 10^{13} for phage displayed polypeptides, and into the millions for libraries of small molecules. These libraries are then screened or subjected to selection in order to find useful candidate compounds.

Typical screening procedures, as specified, e.g., in U.S. Patent No. 5,824,514 to S. Kauffman and Balivet, incorporated herein by reference in its entirety, are based on the use, for example, of a ligand as the screen, and then screening for a novel molecule able to bind the ligand. For example, the ligand might be the estrogen receptor and phage display libraries are searched for novel peptides or polypeptides able to bind the estrogen receptor. Any such peptide or polypeptide is a candidate drug which might mimic, modulate, agonize, or antagonize the action of estrogen. In an equivalent procedure, the SELEX procedure, an RNA molecule able to bind a target is selected. Typically, once an initial set of candidate compounds is located, the candidates are in one form or another, "amplified" or replicated and then subjected to successive binding and amplification cycles in order to winnow down to good binding candidates. The U.S. Patent No. 5,824,514, Patent No. WO9424314 to S. Kauffman and J. Rebek, S. Brenner et al. Proc. Natl. Acad. Sci. USA, 1992, 89:5381, are

hereby incorporated in their entireties as non limiting examples of generating, characterizing, and screening molecular diversity libraries.

The term in the art for the early stages of the drug development process is called "lead discovery," and a successful candidate is referred to as a "lead."

Three gaps in the current technology are becoming increasingly important to close:

1) Consider screening a population of molecules for candidates that mimic the shape and/or structure of some target compound. Present screening or selection techniques primarily identify candidates that are very close in shape and/or function to the target compound, *e.g.*, candidates that efficiently bind to receptors or antibodies of the target compound. Therefore, molecules that are "close" in shape and/or structure to a viable mimic for the target molecule, but not similar enough to bind efficiently a target receptor or target antibody remain undetected by current screening procedures.

2) A second, pressing problem in the drug discovery field is referred to herein as "The Multiple Target Problem". Typical drug compounds must satisfy a number of criteria. For example, a compound may be sought that selectively binds a particular receptor in preference to one or more different receptors, crosses cell membranes and nuclear membranes, survives oral ingestion, does not cross the blood brain barrier, shows good renal clearance, and does not exhibit a variety of cross binding properties to other molecules or sites that would cause further side effects. The process of further developing or modifying a lead compound to meet such criteria is often defined as lead optimization. Lead optimization is financially and labor intensive. For example, if the total cost of development of a drug, including clinical trials, is on the order of 200 to 300 million dollars, then the typical cost of lead discovery might be on the order of 1 million dollars, while lead optimization may typically cost 20 to 40 million dollars. That is, lead optimization is a far more expensive and complex step in the drug development process than lead discovery. Indeed, the very field of molecular diversity is making the discovery of good leads ever easier, hence commoditizing the discovery of drug leads.

Problems 1 and 2 above are related: Solving the multiple target problem, in general, requires finding a set of initial candidate molecules or leads able to meet a number of different criteria. One would expect to find initial candidates that were only slightly able to perform several or all of the tasks, then optimizing one or more candidates until optimum (perhaps compromise) candidates were obtained. Thus, the capacity to find candidates to fulfill multiple tasks simultaneously is typically going to require the ability to locate and optimize molecules which are initially quite poor at all or most of the tasks.

3) In order to solve the multiple target problem, it will be necessary to generate ever "improved" libraries of candidate molecules "spotted into" the proper region of

molecular shape space. Thus, suppose one wished to find a molecule able to bind the estrogen receptor and also able to bind some other receptor, X. Initial candidates might be poor at both tasks, so poor that one could not obtain binding either to the X or the estrogen receptor. It is an object of the present invention to detect molecules that are modestly close to being able accomplish both tasks, i.e., bind both the estrogen and X receptors. Then, the initial screen will have identified a good region of shape space where candidates to solve both tasks are located. Then, further screening or selection would be enabled by the capacity to generate a new library of candidate molecules in the vicinity of this good region of shape space and select or screen for candidates with improved capacities to accomplish both tasks. A succession of such steps, generating and testing new libraries directly or in part computationally, would then constitute a lead optimization procedure with respect to these two tasks.

In practice, traditional applications of combinatorial chemistry and or molecular diversity have faced an additional obstacle: with respect to combinatorial chemistry, it has proved difficult to deconvolve highly diverse libraries of small organic molecules, and pharmaceutical companies are moving in the direction of attempting to make focused libraries, typically built by derivatization of a common core molecular structure at many sites, each in many ways. In short, screening is tending to move from high throughput - high diversity libraries, to restricted or low diversity libraries, even batches of ten or fewer molecular species. Thus, there is a need in the art for enhanced means to generate focused libraries of high diversity, but localized to specific regions of shape space.

Prior to the onset of molecular diversity, combinatorial chemistry, and high throughput screening, rational drug design was the procedure of choice to construct one or more molecular species to test as potential drugs. Here the aim has been to understand the target "receptor" site, such as, for example, the structure, conformation, pose, or epitope of the site and rationally design the candidate drugs to bind to the target site.

Without the guidance of initial candidate molecules to assist the rational design process, however, rational drug design can be a labor intensive time consuming process, which may not readily arrive at an ideal drug design. Thus, there is a need in the art to find a means to marry molecular diversity - combinatorial chemistry with rational drug design to overcome the problems above, such that initial and optimized leads can be achieved at lower cost and higher efficiency. The present invention achieves these needs.

3. SUMMARY OF THE INVENTION

The present invention preferably provides a method for predicting a property of a molecule comprising the steps of:

obtaining an initial odd set of molecules that bind at least one molecule belonging to an origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

selecting a training set comprising a subset of the even set of molecules;

determining a conformation for each of the training set molecules

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subset of the even set of molecules and wherein the new molecule has a new conformation and the model includes the conformation of at least some of the training set molecules; and

predicting the predetermined property of the new molecule.

In a preferred embodiment, the method further comprises selecting a training set comprising a subset from the odd set of molecules and repeating the determining, constructing, and predicting steps wherein the model further comprises the conformation for each molecule in the odd subset.

In another embodiment, the method comprises predicting a predetermined property of at least one molecule assigned to one of the subsets, conditionally modifying the model in response to a difference between said predicted predetermined property and an empirical estimate of said predicted property, and repeating said predicting and conditionally modifying steps until said difference reaches a predetermined value.

In one embodiment, the predetermined property is the ability to bind to at least one predetermined molecule.

The model may comprise at least one of a neural network, a factor analysis, or a principal components analysis.

The neural network may comprise a plurality of layers each having at least one node wherein the plurality of layers include a first layer having at least one node coupled to an input value and a second layer having at least one node coupled to a plurality of nodes of said first layer and a first layer having at least one node with a first transfer function and the second layer having at least one node with a second transfer function.

The method of claim conformation is determined by at least one of x-ray crystallography, nuclear magnetic resonance, or ab initio molecular modeling.

The conformation may comprise at least one of an absolute positions of atomic nuclei in each molecule, a relative position of atomic nuclei in each molecule, an electron density distribution, a bond angle, a bond length, or a van der Waals radii of atoms in the molecule. A conformational data base may be searched for molecules having a conformation similar to the new conformation. At least a portion of the new molecule may be synthesized.

The new molecule preferably comprises at least one of DNA, RNA, a peptide, a polypeptide, or a small molecule. In another embodiment, one or more first variants of the new molecule that are at least somewhat similar to the new molecule are produced and one or more of the first variants having at least one desired characteristic are selected. The first variants preferably comprise a stochastic sequence of polynucleotides.

In another embodiment, antibodies are raised against the new molecule.

Another embodiment of the present invention comprises the steps of:

obtaining an initial odd set of molecules that bind at least one molecule belonging to an origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

obtaining an odd set of molecules that bind at least one molecule belonging to said even set of molecules;

repeating said obtaining an odd set of molecules and said obtaining an even set of molecules steps to generate a sequence of odd and even sets of molecules wherein the molecules in each of said sets bind to at least one of the molecules in a preceding one of the sets in the sequence; and

selecting a training set comprising an even subset from each of at least two even sets of molecules;

determining a conformation for each molecule in each of said subsets;

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subsets of molecules wherein the model comprises the conformation of at least some of the molecules from each even subset; and

predicting a predetermined property of the new molecule.

In a preferred embodiment, the method further comprises selecting a training set comprising a subset from each of at least two odd sets of molecules and repeating the determining, constructing, and predicting steps wherein the model further comprises the conformation for each molecule in each of odd subsets.

In a more preferred embodiment, the method further comprises the steps of:

predicting a predetermined property of at least one molecule assigned to one of the subsets;

conditionally modifying the model in response to a difference between said predicted predetermined property and an empirical estimate of said predicted property; and

repeating said predicting and conditionally modifying steps until said difference reaches a predetermined value.

The predetermined property is preferably the ability to bind to at least one predetermined molecule.

The model preferably comprises at least one of a neural network, a factor analysis model, a principal components analysis model, or an independent component analysis model.

Another embodiment of the invention comprises the steps of:

selecting a first origin set of molecules;

obtaining an initial odd set of molecules that binds at least one molecule belonging to the first origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

obtaining an odd set of molecules that bind at least one molecule belonging to said even set of molecules;

repeating said obtaining an odd set of molecules and said obtaining an even set of molecules steps to generate a sequence of odd and even sets of molecules wherein the molecules in each of said sets bind to at least one of the molecules in a preceding one of the sets in the sequence;

selecting a second origin set of molecules and repeating said obtaining an initial odd set, obtaining an even set, obtaining an odd set and said repeating steps to generate a second sequence of odd and even sets of molecules;

selecting a training set comprising an even subset from each of at least two even sets of molecules belonging to the first and second sequences;

determining a conformation for each molecule in each of said subsets;

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subsets of molecules wherein the model comprises the conformation of at least some of the molecules from each even subset; and

predicting a predetermined property of the new molecule.

Preferably, the predetermined property comprises the ability of the new molecule to bind to each of at least two predetermined molecules.

4. BRIEF DESCRIPTION OF THE FIGURES

FIG.1 is a flowchart representing an example of a process for obtaining a targetshape group.

FIG. 2 discloses a representative computer system in conjunction with which the embodiments of the present invention may be implemented.

5. DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The present invention has as its object, a means to obtain one or more initial candidate molecules, e.g., lead molecules in a drug discovery process, that are at least somewhat dissimilar to a chosen target molecule or "targetshape", to produce initial variants of the initial candidate molecules, and to screen or select candidates from among those variants for molecules that are either more or less similar to the targetshape. The process of producing variants and screening or selecting from among the variants for molecules may by repeated at least once. Hence, the present invention provides a means to carry out an adaptive walk in molecular shape space to climb towards or away from close mimics of a given targetshape.

A further object of the present invention is to provide a means of generating diversity libraries of candidate molecules that are "focused" into a selected region of shape space. For example, without limitation, consider the problem of finding a molecule able to bind the estrogen receptor and also able to bind some other receptor, X. Initial candidates might bind so weakly to both receptors as to be undetectable. The present invention provides a means of identifying initial candidate molecules that are only modestly "close" to being able to bind both the estrogen and X receptors. These initial candidates likely occupy the same region of molecular shape space that is or would be occupied by improved candidates that are better able to bind both receptors. Improved candidates can be sought by generating a new library of candidate molecules focused in the same general vicinity of shape space as the initial candidates and selecting or screening for improved candidates better able to bind simultaneously both receptors. An iterative succession of such steps, obtaining and testing new libraries directly or in part computationally, constitutes a lead optimization procedure with respect to these two tasks.

A variety of characteristics may be used to select molecules according to the invention. According to one mode of carrying out the process according to the invention, the property serving as the criterion of selection is that of having at least one epitope

similar to one of the epitopes of a given antigen or other molecule. According to another mode of the invention the criterion for selection may be the capacity of a molecule to bind a given antigen, other molecule, or surface. According to yet another mode, the criterion for selection may be the capacity of a molecule to displace a member of two or more bound molecules.

The property serving as the criterion for selection can be the capacity of the molecule to catalyze a given chemical reaction. For instance, for the production of several peptides and/or polypeptides, the said property can be the capacity to catalyze a sequence of reactions leading from an initial group of chemical compounds to at least one target compound.

The said property can also be the capacity to modify selectively the biological or chemical properties of a given compound, for example, the capacity to selectively modify the catalytic activity of a polypeptide or other molecular catalyst.

The said property can also be the capacity to stimulate, inhibit, or otherwise modify at least one biological function of at least one biologically active compound, chosen, for example, among the hormones, neurotransmitters, adhesion factors, growth factors, and specific regulators of DNA replication and/or transcription and/or translation of RNA.

The invention also has as its object the use of the molecule obtained by the processes of the invention, for the measurement, e.g., qualitatively, quantitatively, or both of an analyte or other target molecule.

According to a particularly advantageous mode of carrying out the invention, the desired characteristic of the molecule is the capacity to simulate or modify the effects of a biologically active molecule, for example, a protein, and screening and/or selection for clones of transformed host cells producing at least one peptide or polypeptide having this property, is carried out by preparing antibodies against the active molecule, then utilizing these antibodies after their purification, to identify the clones containing this peptide or polypeptide, then by cultivating the clones thus identified, separating and purifying the peptide or polypeptide produced by these clones, and finally by submitting the peptide or polypeptide to an in vitro assay to verify that it has the capacity to simulate or modify the effects of the said molecule.

It is known in the art that, as a non-limiting example, if one screens a phage display or RNA aptomer library for sequences that bind to a receptor, a number of different sequences will do so with a distribution of affinities. If one takes the set of sequences binding the receptor above a chosen threshold of affinity, and examines the sequences, it is known in the art that this set can often be organized into one or more families, each of which has a consensus sequence. Thus, the consensus sequence itself, plus some family of

related sequences, are candidates to bind the receptor. Often, but not always, the consensus sequence itself, if constructed, will bind the receptor.

The invention carries over to obtaining polypeptides by the process specified above and utilizable as chemotherapeutically active substances.

5.1 Molecular Diversity Libraries

5.1.1 Recombinant Techniques

A variety of means are available for the generation of molecular diversity libraries. For example, and not by way of limitation, a process for obtaining DNA, RNA, peptides, polypeptides, or proteins through the use of transformed host cells containing genes capable of expressing these RNA's peptides, polypeptides, or proteins, i.e., by recombinant DNA techniques as described in U. S. Patent No. 5,824,514 to S. Kauffman et al., U.S. Patent No. 5,763,192 to S. Kauffman et al., U.S. Patent No. 5,723,323 to S. Kauffman et al., M. Pavia et al. *Bioorg. & Med. Chem. Ltrs.*, 1993, 3:387, and J. Devlin, et al. *Science*, 1990, 249:404 which are hereby incorporated by reference in their entireties. Using such techniques, a library of expression vectors containing stochastically generated polynucleotide sequences is formed. Host cells containing the vectors are cultured so as to produce peptides, polypeptides, or proteins encoded by the stochastically generated polynucleotide sequences. Screening or selection is carried out on such host cells to identify a peptide, polypeptide or protein produced by the host cells which has a predetermined property. The stochastically generated polynucleotide sequence which encodes the identified peptide, polypeptide, or protein is then isolated and used to produce the peptide, polypeptide, or protein have the predetermined property.

5.1.2 Random Chemistry

Another approach to generating a diversity of compounds is described in Patent No. WO9424314 to Kauffman and Rebek, incorporated herein by reference in its entirety, discloses the generation of new compounds using random chemistry, with or without enzymes, and the subsequent characterization or identification of compounds with a desired property.

In one random chemistry approach, a starting group of different organic molecules is provided. At least one chemical reaction is caused to take place with at least some of the different organic molecules in the starting group to create an intermediate reaction mixture having one or more organic molecules different from the organic molecules in the starting group. The step of causing at least one chemical reaction to take place is repeated at least once. Subsequent repetitions uses the reaction mixture of the previous step, and in the end

produces a final reaction mixture as a result of the last repetition. The final reaction mixture is screened for the presence of the organic molecule having a desired property.

In another approach to random chemistry, a diversity of compounds is generated from a group of substrates which are subjected to a group of enzymes representing a diversity of catalytic activities. As used herein, the term "enzyme" includes enzymes (e.g., naturally or non-naturally occurring or produced), catalysts (e.g., catalytic surfaces), candidate catalysts and candidate enzymes (e.g., antibodies, RNA, DNA or random peptides/polypeptides). The substrates may have different or similar core structures, and similar or different functional groups as substituents. Alternatively, the substrates may have different or similar core structures and different or similar functional groups as substituents. The substrates may have similar or identical core structures, but a variety of different functional groups as substituents permitting the creation of a diversity of compounds centered around a particular compound or a particular class of compounds.

For example, one may react a group of different enzymes representing a diversity of catalytic activities under suitable conditions with a group of different substrates, thereby producing one or more organic molecules different from the enzymes and substrates in the reaction mixture; screen the reaction mixture for the presence of an organic molecule having a desired property; and isolate from the reaction mixture the organic molecule having the desired property. In another approach, one may react a group of different enzymes representing a diversity of catalytic activities under suitable conditions with a group of different substrates, thereby producing one or more organic molecules different from enzymes and substrates in the reaction mixture; screen the reaction mixture for the presence of an organic molecule having a desired property; and determine the structure or functional properties characterizing the organic molecule have the desired property.

Using a random chemistry approach, at least two ways are provided for generating a diversity of molecules, one which does not use enzymes, but uses a variety of possible adducts or other molecules which may undergo reactions with the initial molecule of interest, and also uses a variety of chemical reagents and physical conditions to drive the synthesis of a library of derivatized products of the initial molecule. Alternatively, the core initial molecule plus a set of candidate adducts and other molecules which may react with the initial molecule are used, but also included is a set of enzymes which may increase the rate of formation of the local high diversity library of derivatized forms of the initial compound. It will be readily appreciated by those of ordinary skill in the art that the methods for producing general high diversity libraries of product molecules and for producing local high diversity libraries of derivatized forms of an initial compound may be

combined. For example, a new initial compound may be generated by the general procedure (e.g., substrates with different core structures). Such a new compound is then used, with or without derivatives, to generate a local high diversity library of derivatized forms of the compound. Further, it will be evident to those of ordinary skill in the art that libraries may be generated using a combination of random chemistry methods without enzymes and with enzymes.

5.1.3 Production of Antibodies

Described herein are methods for the production of antibodies capable of specifically recognizing one or more target epitopes or molecules. Such antibodies may include, but are not limited to polyclonal antibodies, monoclonal antibodies (mAbs), humanized or chimeric antibodies, single chain antibodies, Fab fragments, $F(ab')_2$ fragments, fragments produced by a Fab expression library, anti-idiotypic (anti-Id) antibodies, and epitope-binding fragments of any of the above. Such antibodies are useful as shape complements to one or more target molecules as part of a molecular diversity library according to the invention.

For the production of antibodies to target epitope or molecule, various host animals may be immunized by injection with the target molecule a portion thereof. Such host animals may include but are not limited to rabbits, mice, and rats, to name but a few. Various adjuvants may be used to increase the immunological response, depending on the host species, including but not limited to Freund's (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, dinitrophenol, and potentially useful human adjuvants such as BCG (bacille Calmette-Guerin) and *Corynebacterium parvum*.

Polyclonal antibodies are heterogeneous populations of antibody molecules derived from the sera of animals immunized with an antigen, such as target molecule, or an antigenic functional derivative thereof. For the production of polyclonal antibodies, host animals such as those described above, may be immunized by injection with the target molecule supplemented with adjuvants as also described above.

Monoclonal antibodies, which are homogeneous populations of antibodies to a particular antigen, may be obtained by any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to the hybridoma technique of Kohler and Milstein, (1975, Nature 256:495-497; and U.S. Patent No. 4,376,110), the human B-cell hybridoma technique (Kosbor et al., 1983,

Immunology Today 4:72; Cole et al., 1983, Proc. Natl. Acad. Sci. USA 80:2026-2030), and the EBV-hybridoma technique (Cole et al., 1985, Monoclonal Antibodies And Cancer Therapy, Alan R. Liss, Inc., pp. 77-96). Such antibodies may be of any immunoglobulin class including IgG, IgM, IgE, IgA, IgD and any subclass thereof. The hybridoma producing the mAb of this invention may be cultivated in vitro or in vivo. Production of high titers of mAbs in vivo makes this the presently preferred method of production.

In addition, techniques developed for the production of "chimeric antibodies" (Morrison et al., 1984, Proc. Natl. Acad. Sci., 81:6851-6855; Neuberger et al., 1984, Nature, 312:604-608; Takeda et al., 1985, Nature, 314:452-454) by splicing the genes from a mouse antibody molecule of appropriate antigen specificity together with genes from a human antibody molecule of appropriate biological activity can be used. A chimeric antibody is a molecule in which different portions are derived from different animal species, such as those having a variable region derived from a murine mAb and a human immunoglobulin constant region.

Alternatively, techniques described for the production of single chain antibodies (U.S. Patent 4,946,778; Bird, 1988, Science 242:423-426; Huston et al., 1988, Proc. Natl. Acad. Sci. USA 85:5879-5883; and Ward et al., 1989, Nature 334:544-546) can be adapted to produce antibodies to one or more target molecules. Single chain antibodies are formed by linking the heavy and light chain fragments of the Fv region via an amino acid bridge, resulting in a single chain polypeptide.

Antibody fragments which recognize specific epitopes may be generated by known techniques. For example, such fragments include but are not limited to: the $F(ab')_2$ fragments which can be produced by pepsin digestion of the antibody molecule and the Fab fragments which can be generated by reducing the disulfide bridges of the $F(ab')_2$ fragments. Alternatively, Fab expression libraries may be constructed (Huse et al., 1989, Science, 246:1275-1281) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity.

5.2 Detection of Molecules and Molecular Binding

A variety of means are available which allow characterization, e.g., measurement quantitatively, qualitatively, or both, of low concentrations of one or more species of a desired molecule in a mixture of molecules generated by the methods provided herein. A variety of means are also available which allow characterization of binding or affinity between molecules.

In general, the methods of the invention comprise ascertaining the presence of a molecule having a desired property and/or measuring the abundance of a molecule having a desired property in a set or mixture of molecules generated by the methods provided herein.

A variety of cell systems are well known to those of ordinary skill in the art which allow measurement of low concentrations of ligands, e.g., ligands binding a hormone receptor. In this regard, for example, a system has been developed which clones human G peptide hormone receptors into frog melanocytes (Lerner, Proc. Natl. Acad. Sci. USA).

The hormone receptors, typically located in the cell membrane, respond to binding of the corresponding hormone, but trigger a cell response releasing or reabsorbing melanophores. In a forty minute reversible cycle, cells darken dramatically, then can be induced to lighten in color again. Response of the cell depends upon the affinity of the hormone for the receptor. Typical responses occur in the nanomolar to 100 picomolar hormone concentration range. For some hormone receptorhormone pairs where affinity is higher, response occurs in the picomolar hormone concentration range. This cell system is an example of an assay system which allows measurement, in a mixture of molecules, of one or more species of ligands able to bind to the receptor. The set of molecule ligands able to bind the receptor are then the ligands of interest, for they are candidates to act as drugs by antagonizing, agonizing, substituting for, or modifying the effects of the natural hormone. Alternatively, according to the methods of the invention, the ligands of interest may be those not binding the receptor.

A second example of a cell assay is that available commercially from Molecular Devices (Palo Alto, CA). It consists of an array of chemfets which respond to very small changes in local pH. In turn, these small pH changes reflect the altered metabolic activity of a population of cells upon receipt of some molecular signal, such as a hormone binding its receptor. For example, cell assays in which a hormone binds a receptor are known to those of ordinary skill in the art and allow nanomolar or subnanomolar concentrations of the hormone ligand to be measured. A preferred means of using the present invention consists in exposing such cells to a high diversity library of molecules or target shape set of molecules generated by the methods provided herein1 to ascertain the presence of or measure the abundance of one or more species of molecules able to trigger the cell response. That set of molecules, each of which is highly likely to bind the hormone receptor are the molecules of interest.

Another example is to use blast B cells, which on their surface express antibodies directed to a molecule of interest, to detect in a high diversity library the presence of molecules which sufficiently mimic the molecule of interest to be able to bind to its

antibody on a B cell. Thus, an animal is immunized with a molecule of interest and the early B cells isolated. A high diversity library of molecules generated by the methods provided herein is screened using the population of B cells. For example, binding may stimulate cell cycling or division by the last B cell bound. Cell cycling or division may be detected by means known in the art.

Alternatively, a variety of assays to detect the presence of a ligand of interest exist which are based on direct binding assays. Thus, for example, a receptor for a hormone can be used directly to detect binding of a radioactivity labeled ligand. Other means, known in the art, to accomplish this include the following:

(i) The estrogen receptor is used as a non-limiting example. The cloned receptor can be affixed to a flat surface, for example, a filter. Very high specific activity estrogen is prepared, and bound to the receptor population. This set of bound receptors is then used in a competitive assay. The bound receptors are exposed to a library of compounds generated by the methods of the present invention. If the library contains ligands which also bind the estrogen receptor, those ligands will compete with the radioactively labeled estrogen itself for the receptors. Hence the radioactively labeled estrogen will be competitively displaced from the receptors and can readily be detected by means known in the art. Thus, this assay allows detection of one or more species of ligands in the mixture which compete with estrogen for the estrogen receptor. This set of ligands is the set of interest, as they are candidates to be drugs mimicking or antagonizing estrogen.

(ii) The estrogen receptor is again used as a nonlimiting example. By means known in the art, one raises antibody molecules which are able to bind the receptor when the receptor is not bound by estrogen, but not bind the receptor when occupied by estrogen. Alternatively, one generates antibody molecules which bind the estrogen receptor only when the receptor itself does bind estrogen. These antibody molecules can then be decorated with reporter groups by a variety of means known in the art, and used to detect the presence of one or more ligand species in a library of high diversity, which bind to the estrogen receptor. In the case of antibodies which only bind the receptor if the receptor is itself unbound by estrogens, one tests for loss of antibody binding in the presence of the library of compounds and in the simultaneous absence of estrogen. In the case of antibodies which bind the receptor only if the receptor is bound by estrogen, one tests for an increase in binding of the antibody in the presence of the receptor and high diversity library.

(iii) In order to detect ligands in a high diversity library which are candidates to mimic or antagonize the action of a given hormone or other molecule of interest, it is advantageous to generate one or more monoclonal antibodies which bind the hormone or

other molecule of interest. This set of monoclonal antibodies can then be used, rather than a receptor, for the target molecule that is to be mimicked, in binding assays such as those noted above to detect the presence of one or more ligand species in the reaction mixture which are candidates to mimic or antagonize the action of the target molecule. An advantage of this procedure is that a receptor for the target molecule need not be available. Use of a set of monoclonal antibodies is advantageous because, a priori, it is not certain which molecular feature, or epitope, of the target molecule mediates its biological action. Use of a set of monoclonal antibodies, each responding to a different epitope on the target molecule, enhances the probability that the ligands detected in the high diversity library will include those which mimic the biologically important epitope of the target. In some cases it may be possible to selectively use only those monoclonal antibody molecules which bind to the known important epitope of the target molecule.

(iv) Means are established in the art to measure protein-protein binding based on plasmon resonance and detection of a shift in refractive index. In a detection system developed by Pharmacia (Piscataway, NJ), a monoclonal antibody, or a hormone receptor is layered onto a gold chip. Binding of hormone, or other ligands to a receptor, is measured in very low concentrations (e.g., in the nanogram range or less). Thus, any receptor, or antibody, or other "shape complement" of a target molecule of interest can be placed on the gold chip, the latter can be exposed to a high diversity library, and the presence of binding species can be measured quantitatively, qualitatively, or both.

Another example of direct measurement of ligand-binding, which the applicant believe was developed by Evotech, can measure ligand binding in the femtomolar range.

A variety of approaches for characterizing molecular binding are based on fluorescence correlation spectroscopy. For example, Rudolph Rigler 1995, J. Biotechnology, 41:177 has reviewed fluorescence correlation approaches to measuring molecules and binding of molecules. In one approach, a laser beam is focused to a radius of less than about 1 micron. Fluorescent molecules or molecules labeled with fluorescent tags can be measured at femtomolar concentrations, (10^{-15} M), in tens of seconds. Binding of a fluorescent molecule or a molecule labeled with a fluorophore to another molecule can be characterized, e.g., measured qualitatively, quantitatively or both because of the reduced diffusion coefficient of the bound molecules compared to the unbound molecules. Similar approaches based on the different electrophoretic mobilities of bound and unbound molecules are known in the art. Competitive assays in which a molecule displaces a member of at least two bound molecules can be used to assess the relative binding efficiency or affinity of a set of molecules for one or more other molecules.

Thus, for example, if estrogen is a target molecule, and a small RNA aptomer is a shape complement which binds estrogen, then fluorescently labeled versions of that RNA aptomer can be used in a fluorescence correlation approach. An estrogen-mimic which binds the fluorescently labeled RNA will slow its diffusion as detected in the laser system. Thus estrogen-mimics at very low, 10^{-15} M or femtomolar, concentrations can be detected. Alternatively, one may begin with a number of complexes comprising estrogen and the fluorescently labeled RNA aptomer. Adding one or more molecules that compete with the RNA aptomer for binding sites on estrogen can be detected by the appearance of unbound labeled RNA aptomer. One or ordinary skill in the art recognizes that a number of possible approaches for detecting the binding and binding characteristics using a fluorescence based approaches are possible.

A further means to detect ligands of interest at very low concentrations consists in seeking ligands which block a DNA polymerase. By blocking the DNA polymerase chain reaction (PCR) enzyme, amplification of the DNA can be blocked. Since PCR amplification can yield billions or more copies of the initial DNA sequence, blocking PCR amplification yields a readily detectable signal of a ligand which blocks the polymerase. Clearly, this method generalizes to other means to amplify DNA, RNA, or DNA- or RNA-like molecules such as ligation amplification, and extends to general means to block polymerases directly or indirectly with ligands of interest.

As described herein, compounds of interest may act as catalysts for a desired reaction, or as cofactors with other molecules to form an active catalyst. Other molecules may act as inhibitors of enzymes. In order to exclude the possibility that the enzymes or catalysts are found among the candidate set of enzymes which may have been used to generate the compounds of interest, the latter set of enzymes can be quantitatively removed from the high diversity library by, for example, affinity columns bearing molecules directed to a constant part of each of the set of enzymes, or other means known in the art. The resulting high diversity library itself is then assayed for candidates of interest.

Detection of molecules able to inhibit an enzyme may proceed by detecting ligands able to bind the enzyme, as described above. Identifying molecules which are candidates to catalyze a reaction alone or as a cofactor may proceed by testing high diversity libraries of the invention alone, or in the presence of a helper molecule, say a protein, for which a desired molecule will be a cofactor. The system is tested for the presence of ligands able to bind a stable analogue of the transition state of the reaction. Such binding molecules are the candidate catalysts or cofactors sought, for they are candidates to catalyze the reaction itself.

Alternatively, a variety of means are known in the art which allow detection of the products of a catalyzed reaction itself. For example, chromogenic or fluorogenic substrates for a variety of reactions of interest are available. Catalysis of the reaction increases the rate of formation of the colored or fluorescent product. Alternatively, assay systems are available or readily prepared which detect the presence of a product molecule because that product molecule binds a receptor an antibody molecule, or other shape complement. Thus, detection of higher rates of formation of that product molecule demonstrates that the reaction itself was catalyzed.

5.3 Characterization of Molecular Libraries

Following the generation of high diversity libraries of compounds and the screening for the presence of compounds having properties of interest, compounds of interest may be characterized with or without isolation. A variety of means, including those known in the art, are available to characterize or isolate such compounds of interest.

Characterization and/or isolation depend upon the information desired and can be carried out at different mole abundances of the target molecule of interest. For example, using modern mass spectrographic analysis, about 10^{-15} to 10^{-18} moles can be assayed for mass and charge, then fragmented in a variety of ways known in the art and the fragments assayed for mass and charge. Using such data, it is possible to derive the structure of the molecule of interest. For example, ligands of interest may be isolated by binding to a given hormone receptor, or monoclonal antibody, then the liganding molecules released by means known in the art and finally characterized analytically. One means comprises attaching a target receptor or antibody to a solid support. A reaction mixture or subset thereof is contacted with the solid support. Those molecules that are bound will be retained, while the non-bound molecules are readily separated from the solid support. The molecules of unknown structure which have been retained, are then eluted. The freed molecules are characterized analytically, e.g., by mass spectroscopy, NMR, IR, UV, and may be synthesized in batch quantities.

Kibbey et al. U.S. Patent No. 5,670,054, disclose an automated method of sample identification, purification and quantitation wherein a first HPLC column with defined operating parameters is used to separate a small portion of an impure mixture into its constituent components; the individual components corresponding to the eluting zones of the separated mixture are characterized by mass spectrometry; the chromatographic and mass spectroscopic data generated are stored in digital format, for example one compatible with commercial chromatography software, and the data is used to guide the purification of

the remaining sample; the remaining sample is injected on a semi-preparative, or preparative HPLC column; an analog detector output of the semi-preparative, or preparative HPLC system is digitized and evaluated electronically with the previously generated chromatographic and mass spectroscopic data; when elution of a sample component peak corresponding to a desired product peak is sensed, a mechanically actuated, liquid switching value (i.e., a pneumatic or electronic switching valve) is actuated to divert the column eluate from waste to a fraction collection device; and when the end of product peak elution is sensed, the switching valve is actuated to divert the column eluate back to waste collection. The system enables rapid purification of samples in quantities useful for screening of diversity libraries while involving minimal operator input and minimum fraction collection equipment.

In other cases, the concentrations of molecules of interest in the high diversity library will allow detection of their presence, but may be too low for further isolation or characterization. A preferred procedure called "sib selection" allows ready winnowing of the set of candidate enzymes, the set of founder substrates, and the set of reaction conditions and chemical reagents to smaller sets. This winnowing simultaneously reduces the side products generated in the high diversity library, increases the concentration of the target molecule of interest, and identifies the subset of candidate enzymes which catalyze the pathway leading to synthesis of the target molecule, and identifies the set of founder substrates required for synthesis of the desired target. Thus, this sib selection procedure is a means to generate a previously unknown molecule of interest, as well as identify both that molecule and the substrates and enzymes needed to form that molecule.

5.4 Targetshape Groups

A targetshape of molecules, e.g., molecular diversity library, according to the present invention comprises a group of n sets of molecules si (i = 0, ..., n), wherein each set si contains at least 1 molecule. Set s0 generally contains one compound and represents the center or "targetshape" of the group of n sets of compounds. The group of n sets of molecules corresponding to set s0 and having set s0 at its center or origin is referred to as targetshape s. Each set of molecules may also be referred to as a "ring" or "shell." A given ring si is said to have a higher order than ring s(i-1).

Initially, to obtain or generate members of targetshape set s1, a molecular diversity library composed of DNA, RNA, peptides, polypeptides, small molecules, or other compounds is generated or obtained and screened to obtain a set of molecules s1 able to bind a predetermined targetshape or set s0. Alternatively, or in addition, members of s1

may be found by using members of targetshape set s0 to raise antibodies against s0. Given that the members of set s1 bind members of s0, members of s1 generally have at least one epitope or shape feature that is at least somewhat complementary to at least one epitope or shape feature of members of s0.

Members of targetshape set s2 may be found by generating or obtaining and screening a molecular diversity library for molecules able to bind s2 and/or members of s2 may be found by using members of targetshape s1 to raise antibodies against s1.

Typically, for targetshape sets of order $i \geq 2$, any given set si includes a subset of members, si', each of which bind at least one member of set s(i-1) by way of substantially the same epitope or molecular shape feature as those members of set s(i-1) bind at least one member of set s(i-2). Competitive binding assays may be used to identify members of each subset. For example, members of subset s2' may be discriminated from the remainder of set s2 because s2' and s0 will compete for the same binding site on one or members of s1. Given that members of s2' and s0 compete for the same binding site on at least one member of s1, members of s2' generally have at least one molecular epitope or shape feature that is similar to at least one molecular epitope or shape feature of s0. Thus, members of s2' substantially correspond to mimics of s0.

Selecting members of si' by competitive displacement of members of si off s(i-1)' using members of s(i-2)' is analogous to the concept of internal images in the immune system in second rank antiidiotypes and generally corresponds to the search for shape mimics using molecular diversity.

Figure 1 shows an example of a general process for generating or obtaining members of a targetshape group. Initially, an origin set comprised of a targetshape molecule is selected. In the next step, an intermediate set of molecules is generated or obtained. In general, molecules of the intermediate set of molecules bind molecules belonging to the origin set. A terminal set of molecules is then generated or obtained. In general, molecules belonging to the terminal set bind to at least one molecule of the intermediate set. Next, a subset of the terminal set is selected. Generally, molecules belonging to the subset of the terminal set bind at least one member of the intermediate set by way of substantially the same epitope that the one member of the intermediate set binds at least one member of the origin set. In an iterative process, the origin set is replaced with the intermediate set and the intermediate set is replaced with the subset of the terminal set. Finally, the steps of obtaining or generating a terminal set, selecting a subset of the terminal set, replacing the origin set, and replacing the intermediate set can be repeated until a plurality of sets of molecules are obtained.

In general, members of any targetshape set si can be generated or obtained by screening a molecular diversity library composed of DNA, RNA, peptides, polypeptides, small molecules, or other compounds for molecules able to bind at least one member of a lower ordered targetshape subset si-1'. Alternatively, or in addition, members of si may be found by using members of targetshape set s(i-1)' to raise antibodies against s(i-1)'. Random chemistry approaches beginning with molecules having a core structure similar to members of set s(i-2) may also be used generate candidate molecules for set s(i). Subsequently, members of a subset si' that compete for the same binding site on s(i-1)' as members of s(i-2)' may be discriminated from the remainder of set si. In this fashion, members of targetshape sets s0, s1, s2, s2', si, si', ...,sn, sn' may be obtained.

The complete targetshape group of n sets of molecules forms a gradient in shape-function space surrounding the targetshape s0. The even rings i.e., sets of a targetshape group substantially correspond to shape mimics of s0 that are, on average, successively less like s0 as ring order increases, whereas members of a given odd ring substantially correspond to shape complements of molecules belonging to successive even rings. For example, because members of subset s4' are identified by competitive binding with s2' for sites on s3', then, since s2' members are similar to, but not identical to the targetshape s0, it follows that members of s4' are generally more similar to members of s2' than to s0. Consequently, the sets of molecules si, where i is even, comprise a gradient in the shape space surrounding s0 where members of a given subset si' are less similar to s0 than members of the lower ordered subset s(i-2').

By extension, the sets of molecules si, where i is odd, comprise a gradient in shape space surrounding s1 where members of set s(i+2)' are less similar than members of lower ordered subsets si' to the s1 shape complements of s0. It follows that molecules that bind members of odd numbered rings are successively more similar to s0, on average, as they bind to odd numbered rings of lower order. That is, for example, molecules binding members of s5' are generally more similar to s0 than molecules binding members of s7'. Thus, the odd numbered sets provide a complementary shape "gradient" to select or screen for molecules ever more similar to s0.

In selecting members of a given subset si', odd or even, it may be advantageous to proceed by competitive displacement of members of si off s(i-1)' using only members of s(i-2)'. Alternatively, members of set si' may be selected by competition with members of any lower ordered subset s(i-k)' where $i - k \geq 0$ and k is even. Choosing $k > 2$ results in a less-steep gradient because successive subsets are chosen by competition with molecules that are somewhat more similar to either s1 or s0. The value of k need not be the same for

obtaining successive subsets. Thus, the choice of k allows the gradient in shape function space between successive sets of molecules to be tuned at each step in the process.

The steepness of the gradient surrounding s0 or s1 may also be modified by setting more or less stringent competition or binding requirements for entry into a given set or subset.

5.4.1 Multiple Target Problems

The ability to detect lead candidates that do not bind a target efficiently enough to identify by available screening techniques allows the method described above to be extended to address the multiple target problem. Each characteristic that the molecule must posses or criterion that the desired molecule must satisfy can be considered a "task." Without loss of generality, and as an example only, consider the problem of finding a molecule able to accomplish two tasks such as having the ability to bind receptors of both estrogen and progesterone. Defining estrogen as s0, one may obtain a targetshape group of n sets molecules corresponding to targetshape rings around estrogen. Similarly, define progesterone as r0 and obtain sets of molecules corresponding to targetshape rings around r0. The targetshape groups surrounding each targetshape need not contain the same number of sets of molecules.

A molecular diversity library is then screened for molecules that bind at least one member of an odd ring of both targetshape s and targetshape r. Alternatively, or in addition, members of at least one even numbered ring of targetshape s and/or targetshape r may be used to raise antibodies against the even numbered ring. The antibodies are then screened for molecules that bind at least one odd ring of both targetshape s and targetshape r. Consider a molecule X that binds to s3' and also to r5'. By way of example, suppose that X does not bind to s1 nor to r1. Thus, X would remain undetected using conventional screening tests for molecules binding only the equivalents of targetshape sets s1 and r1. Conversely, using the full targetshape groups surrounding estrogen and progesterone candidates that are only somewhat similar to both estrogen and progesterone may be found.

To obtain compounds with improved binding to both the estrogen and progesterone receptors, variants of X are obtained or generated, for example, using a molecular diversity approach or other random chemistry approach. Alternatively, or in addition, molecules in the same region of shape space as is X are generated or obtained. The new population of molecules is screened to identify members that are more similar to both estrogen and progesterone as ranked by the ability to bind members of lower ordered odd numbered rings of targetshape s and/or targetshape r. Thus, a molecule binding at least one member of s3' and r5' is an improvement of X, for it is more similar to progesterone and as similar to

estrogen. A molecule able to bind at least one member of s1 and r3' is better than X in being more similar both to estrogen and to progesterone.

The screening power of this extended targetshape approach is particularly advantageous in situations where one cannot sample molecular shape space with sufficient density to identify immediately candidates, if they exist, that bind both s1 and r1. For example, although a molecular diversity library may contain approximately 10^{15} distinct molecular species, the chance of identifying a potential candidate to mimic both s0 and r0 may be slim. However, using higher order sets of a targetshape group instead of one or a small number of target compounds, a broader region of molecular shape space can be sampled with high density to identify proto-candidates even only somewhat similar to both s0 and r0. Subsequently, those candidates can be used to generate a molecular diversity of shape variants focused around the same general region of shape space as the initial candidates. Thus, about 10^{15} molecules in that "region" of shape space can be created and screened or selected upon for improved variants. The improved variants can then be used to create still further variants, in an attempt to increase similarity to both s0 and r0.

The process described above can be generalized to seek molecules able to fulfill a plurality of arbitrary "Boolean" or logical combinations of "yes" and"no" conditions on different targets or different criteria. For example, one may seek a molecule that binds one hormone receptor but does not bind the receptor of another hormone, or one may seek a molecule that binds a cis acting promoter of one gene but does not bind the cis acting promoter of another gene, etc. In general, n targetshape groups, r, s, t, ..., n would be constructed for each of n tasks. Candidates only partially fulfilling one or more of the tasks may be identified. Improved candidates could be sought by obtaining variants better able to fulfill each task separately, or by obtaining variants better able to fulfill any subset of the n tasks. Then, the initial candidates may be optimized to seek the practically accessible pareto optimal set.

For example, consider a multitask problem that includes obtaining molecules that bind at least one target, and do not bind at least one other target. As a specific example only, consider seeking a molecule that binds the estrogen receptor but does not bind the progesterone receptor. Targetshape groups are constructed around estrogen, s0, and progesterone, r0. Initial candidates are sought that bind to low ordered odd rings of targetshape s and either will not bind any odd ring of targetshape r, or will only bind higher ordered odd rings of targetshape r so that it is unlikely to interact with or bind the progesterone receptor. However, an initial candidate may bind at least one lower ordered odd rings of targetshape r. By successively generating variants of the initial candidates and selecting improved candidates that bind primarily to higher ordered odd rings of targetshape

r, variants of initial candidates are "tuned" away from similarity to progesterone or interaction with the progesterone receptor. The ability of a variant to bind lower ordered odd rings of estrogen may also be considered when selecting improved variants to continue the optimization procedure. Thus, targetshape provides a means to "sculpt" molecules to enhance the ability to bind the receptor of one molecule while decreasing binding to the receptor of a second molecule. In this example, even subtle side effects due to undesirable binding of a drug to the progesterone receptor would be avoided.

The initial candidates X can be partially ordered based upon their binding to rings of one or more of targetshape groups. It may be advantageous to order fully the candidates according to a system that assigns different weight to the ability to accomplish each task. The candidates can also be ordered based on the absolute or relative number of members of a given targetshape ring that bind each candidate and/or the relative strength or efficiency of binding. However, it is not necessary, even to optimize the ability to accomplish multiple tasks to create a full ordering relationship between all candidate molecules.

A pareto optimal set of candidates, X, is defined as a set of molecules having the property that no other molecules exist that are better than the members of the pareto optimal set with respect to at least one "task" and at least as good with respect to the remaining tasks. In the case where only a partial ordering exists, the pareto optimal set constitutes the "end point" of the effort to find good candidates for both tasks. The pareto optimal set is defined for the set of all possible molecules. Thus, in reality, it is impossible to assure that a candidate pareto optimal set, X, is actually pareto optimal.

In the current context, a pareto optimal set of candidates for the tasks of being similar to estrogen and also to progesterone is a set X such that no other molecule exists that is more similar to estrogen, and at least as similar to progesterone; or that is more similar to progesterone and at least as similar to estrogen.

5.5 Rational Drug Design and Target Shape Groups

Drug design has historically involved "discovering" a particular chemical substance that interacts in some way with receptors, e.g., proteins in the living cells of an organism. As proteins are made up of polypeptides, it is not surprising that some effective drugs are also peptides, or are patterned after peptides. Thus, without limitation, the activity of a drug may be couched within the framework of polypeptides. The description, however, is applicable to any agent capable of modifying the activity of a biological molecule having a receptor. Generally, for two peptides to effectively interact with each other, e.g., one as a protein receptor and the other as a drug, it is necessary that the complex three-dimensional shape ("conformation or pose") of one peptide assume a compatible conformation that

allows the two peptides to fit and bind together in a way that produces a desired result. In such instance, the complex shape or conformation of a first peptide has been compared to a "lock", and the corresponding requisite shape or conformation of the receptor as a "key" that unlocks (i.e., produces the desired result within) the first peptide. This "lock-and-key" analogy emphasizes that only a properly conformed key (second peptide or compound patterned thereafter) is able to bind or fit within the lock (first peptide) in order to "unlock" it (produce a desired result) Further, even if the key fits in the lock, it must have the proper composition in order for it to perform its function. That is, the second peptide must contain the right elements in the right spatial arrangement and position in order to properly bind with the first peptide, e.g., receptor protein. The random diversity or combinatorial approach to drug discovery, as described above, does not require direct knowledge of the confirmation of the target compounds or potential leads that bind the target compounds or are members of a target shape group as described above. As part of the present invention, however, discovering or predicting the proper conformation or shape of the key, or second peptide or compound patterned thereafter can assist the drug discovery, as described below.

Most polypeptide structures exhibit several conformations that are stable, some more so than others. The most stable conformations are the most probable. A conformation may change from one stable conformation to another through the application of sufficient energy to cause the change. Given the opportunity to freely move, fold and/or bend, a given polypeptide chain will eventually assume a stable conformation. The most probable conformation that is assumed is the one that would take the most energy to undo. This most probable conformation is referred to herein as the "global minimum". Other stable conformations are less probable, but may readily be assumed, and are referred to herein as a "local minimum" or "local minima". A conformation that represents a local minimum could thus be changed, through application of an external force, to another stable conformation which is either another, different local minimum or the global minimum.

Only by designing the conformation of one peptide to allow it to fit within the conformation of the other peptide and bind thereto will the desired interaction between the two peptides take place. Thus, principal generic steps in the rational drug design process include: (1) identification and determination of the structure of a receptor site or molecule binding the receptor site; (2) use of theoretical principles and experimental data to propose a series of putative compounds that will bind to the receptor sites, the compounds may be synthesized and tested for complementarity with the active site; (3) determination of the structure of receptor/ligand complexes that bind with high efficiency, *i.e.* with low free energies; and (4) iteration of steps 1-3 to further enhance binding.

Rational drug design thus includes knowing or predicting the conformation of a desired protein receptor peptide. Random chemistry, such as the methods of the present invention described above, provide a starting point to rational drug design approaches by identifying molecules that have suitable activities, such as binding efficiencies, toward a target compound. Indeed, the target shape groups of the present invention allow the identification of potential candidates that would not be identified using known techniques. Moreover, by providing a series of diversity libraries, the target shape rings, the present invention provides a data base of molecular shapes, which form a gradient in conformation space around a particular target. Thus, the present invention is ideally suited to combine a random chemistry approach with a rational design approach in identifying and designing putative drug leads.

The even and odd numbered rings contain information about the requisite sequences and conformations (shape) in molecular shape space that are capable of mimicking a receptor such as, for example, estrogen, or are capable of mimicking its complement, for example, the estrogen receptor. The shape information can be obtained by predicting the shape directly from the sequence information, or directly by NMR, x-ray crystallography, neutron scattering, high-resolution mass spectrometry, crystallization, or other experimental techniques to examine either members of S1, S3, S5, or S2, S4, S6 alone, or as bound pairs, say a member of S1 and a member of S2 that bind one another. Higher order interactions, such as three or more bound molecular species can be examined to ascertain wether a combination of more than one compound can modulate the activity of a receptor such as, for example, by modulating the binding of C to a receptor. In this fashion, for example, a protein phosphorylase can be made to modulate the activation of a transcription factor.

A molecule such as estrogen, which binds a receptor, such as the estrogen receptor, also contains information about the shape needed to bind its receptor. The consensus sequence noted above and derived from the members of even member rings of at least one target shape group themselves contain more information about the sequences and shapes able to bind the estrogen receptor than does estrogen itself. But the members of S2 contain less information about the requisite sequences in shape space to mimic estrogen than does SO, *i.e.* the molecule used to form the target shape group. Additionally, the structures of members of S2, S4, S6, ... provide successively less information about the structure and conformation needed to bind or interact with the receptor. Thus, these molecules for a gradient in shape space, which may be used to guide the selection, search or design of additional candidate molecules using data mining techniques as described below.

Similarly, the estrogen receptor, a member of S1, contains some information about sequence and shape required to bind estrogen, but less information than the entire set S1, and still less than the members of the odd sets, S1, S3, S5, ... Thus, structural activity information from both the even and odd groups can be used in rational drug search, selection, and design.

Neural networks, factor analysis, principal component analysis, independent component analysis, singular value decomposition and other mathematical data mining techniques can be trained on a training set to extract relevant sequence or shape information about molecular sequences, structures and shapes needed, for example, to mimic estrogen, or mimic the estrogen receptor. In general, several factors increase the informing power or predictive power of such data mining techniques. First, the averaging effect of larger data sets acts to minimize the effect of spurious shapes, *e.g.* the shapes of molecules which bind a target away from the active site but are erroneously included in the model. Second, as understood in the art, the informing power is increased when the model or data mining technique includes orthogonal information along several different vectors. By including, for example, molecules from more than one target shape ring the present method advantageously takes advantage of the fact that members of, for example, successively higher ordered even targetshape rings are increasingly different in both structure and binding ability relative to the compound S0. These structural differences help guide the model or data mining technique toward those features which are most important in determining an activity or binding efficiency.

Molecules from both even and odd targetshape groups can be included in the data mining approach. This provides further orthogonal information with which to guide the model. Known application of data mining techniques to rational drug design have not made use of such a gradient of molecules or members of complementary sets of molecules such as members of the odd targetshape groups.

Moreover, as discussed above, the targetshape approach to creating molecular diversity can be applied to a multiple target problem by creating more than one targetshape groups centered around more than one molecule, such as, for example, the estrogen and progesterone receptor. Since the members of rings of each targetshape group can be tested for properties such as renal clearance or liver toxicity and because members of even and odd rings of each targetshape group are families of related shapes, these experiments contain information about the requisite molecule structure and shape features need to obtain pareto optimal drug leads with respect to a set of one or more design criteria. Thus, as a non-limiting example, the present invention can be used with the sparse

membership of the rings of each targetshape group together with data mining techniques such as neural networks to predict structures and shapes that constitute the global pareto optimal surface with respect to two or more design criteria.

Thus, it will be clear that the use of members of the even or odd rings, respectively, can enhance the capacity of nerual networks or other data mining procedures to extract the sequence, structure and shape features required for good binding to or modulation of a molecule or receptor such as, for example, estrogen, or to the estrogen receptor. Clearly, estrogen and its receptor are a non-limiting example. We could equally consider DNA binding molecules, cis site specific DNA binding molecules, molecules binding cis-trans complexes, where those molecules are DNA, RNA, proteins, small molecules or other classes of molecular species such as carbohydrates, lipids, or other polymers, or small molecule families such as those now familiar in combinatorial chemistry.

More generally, the above invention can be used to discover experimentally, or predict rationally, combinations of molecular species, say A and B, which together bind and modulate the activity of, for example, a phosphorylase enzyme acting at some point in a cellular signaling cascade, or a transcription complex binding a DNA or RNA cis site regulating transcription, or otherwise modulating any desired biological activity within or between cells, or organisms a process for discovering A and B may comprise, for example, co-crystallization of the species A and B with, for example, an enzyme whose activity they modulate.

5.5.1 Molecular Structure Determination

The molecular structure of a molecule selected from a screening step can be characterized to predict geometric and conformational features, which lead toward a suitable lead compound. Molecular structure information comprises, but is not limited to, for example, the absolute and relative positions of nuclei, the relative electron density distribution, bond angles, bond lengths, van der Waals radii of atoms in the molecule, chirality, and charge. Additionally, such information can be acquired over all or only a portion of a molecule.

Experimental structure determination can be accomplished, for example, using X-ray crystallography and solution-state nuclear magnetic resonance (NMR) (MacArthur et al., 1994, Trends in Biotechnology 12:149-153).

X-ray crystallography depends on the interaction of electron clouds with X-rays to provide information on the location of every heavy atom in a crystal of interest. The accuracy of X-ray crystallography is 0.5-2.0 Å (1 .Å.=10^{-8} cm). Cocrystallization allows the

structure of more than one bound molecule to be determined providing further information about the conformation of the active site and complex. Such information can be used, for example, to design, select, or optimize molecules that act in tandem to modify the activity of a biological molecule.

Approaches for structure determination using NMR, such as, for example, methods described in *Biomolecular NMR Spectroscopy* by Jeremy N.S. Evans, Oxford University Press, New York, 1995 and *Nuclear Magnetic Resonance of Proteins and Nucleic Acids* by Kurt Wüthrich, John Willey and Sons, 1986, which are incorporated in their entireties by reference. NMR, relies upon correlations between nuclear spins resulting from dipole-dipole interactions indirectly mediated by the electron clouds. High-resolution, multidimensional, solution-state NMR techniques are an attractive alternative to crystallography since that they can be applied in situ (i.e., in aqueous environment) to the study of small protein domains (Yu et al., 1994, Cell 76:933-945). Solution-state NMR has been successful at determining the structure of moderate-sized proteins and protein/ligand complexes to a 2 Å. resolution, similar to that of X-ray methods (MacArthur et al., 1994, Trends in Biotechnology 12:149-153; Clore et al., 1994, Protein Science 3:372-390). The structure of a ligand (such as a pharmaceutical lead compound) bound to a protein is most efficiently determined when the protein is uniformly ^{13}C and ^{15}N labeled, and the binding occurs in the slow exchange limit (Clore et al., 1994, Protein Science 3:372-390). In this limit, a bound complex remains together long enough for resonances of the free and bound form of the ligand to be resolved.

One method that avoids some limitations associated with X-ray crystallography and solution-state NMR and has significant advantages, is solid-state NMR, particularly dipolar-dephasing experiments such as rotational echo double resonance (REDOR) (Gullion et al., 1989, Journal of Magnetic Resonance 81:196-200; Gullion et al., 1989, Advances in Magnetic Resonance 13:57-83). Compared with crystallography, solid-state NMR has the advantage that it obtains high-resolution structural information from polycrystalline and disordered materials. This eliminates the need for the formation of highly regular crystals to achieve high resolution diffraction, and eliminates structural perturbations due to crystal packing forces. In contrast to solution-state NMR, which relies upon mutual correlations between nuclei from the indirect dipolar coupling (studied via the Nuclear Overhauser effect, NOE) that fall off as $1/r^6$, solid-state NMR relies upon the direct dipolar coupling, which decreases as $1/r^3$, for the measurement of internuclear distances, where r is the internuclear distance. As a result, longer distances can be measured with solid-state NMR, and the distances measured have a higher degree of accuracy and precision.

Furthermore, solid-state NMR is not strictly limited by the size of the complex resulting from the drug bound to a target molecule. In the solid-state NMR experiments, the size limitations are determined primarily by the quantity of the sample available, and the sensitivity of the NMR spectrometer.

One advantage of REDOR transform technique over solution-state NMR measurement is the direct and accurate determination of the internuclear distance from a measured frequency. Solution-state NMR experiments rely upon the indirect measurement of the dipolar coupling for distance measurements. In solution-state NMR there is no direct relationship between an experimentally measured parameter and the distance. Instead, the strength of the coupling, as inferred from the Nuclear Overhauser effect (NOE), is related to a range of possible distances spanning a few Angstroms.

In addition to REDOR, other dipolar-dephasing (dipolar-recoupling) methods such as TEDOR (Hing et al., 1993, Journal of Magnetic Resonance, Series A 103:151-162; Hing et al., 1992, Journal of Magnetic Resonance 96:205-209), DRAMA (Tycko et al., 1990, Chemical Physics Letters 173:461-465; Tycko et al., 1993, Journal of Chemical Physics 98:932-943), DRAWS (Gregory et al., In 36th Experimental Nuclear Magnetic Resonance Conference; Boston, Mass., 1995; p 289), and MELODRAMA (Sun et al., 1995, Journal of Chemical Physics 102:702-707), are known in the art.

As applied to biological materials, REDOR has primarily been used to determine the distance between one ^{13}C atom and one ^{15}N atom (Marshall et al., 1990, Journal of the American Chemical Society 112:963-966; Garbow et al., 1993, Journal of the American Chemical Society 115:238-244). Because of the nature of nuclear magnetic interactions in the solid state, REDOR has the inherent ability to measure internuclear distances with a high degree of accuracy and precision. REDOR measurements are accurate to better than 0.05 Å when the ^{13}C-^{15}N distances are from 0 to 4 Å, and to better than 0.1 Å when the ^{13}C-^{15}N distances are from 4 to 6Å. Garbow and Gullion (Garbow et al., 1991, Journal of Magnetic Resonance 95:442-445) have shown that data acquisition rate using REDOR can be enhanced by the measurement of REDOR signals from chemically shifted nuclei.

5.5.2 Rational Design of Molecules and Prediction of Activity

The success of rational drug design approaches can be enhanced through the use of computation methods to identify key molecular portions or fragments, which are relevant to the binding or activity of the putative drug leads. In order to predict the activity of molecules not yet synthesized or for which not much is known with respect to a particular chemical function, such as binding to a particular receptor, one would first start with

molecular structures and assay values of known molecules with known activities with respect to such chemical function. Examples of such methods include Rodgers D. and Hopfinger A. J., J. Chem. Inf. Comp. Sci., 1994, 34, 854-866, which is incorporated in its entirety by reference. The molecular structure information determined using the methods described above and assay information may provide input information for a neural network. As understood by one of ordinary skill in the art, neural networks may have a variety of structures and the example described below illustrates but one approach predicting the activity of a new molecule or for predicting pertinent conformational features of the active site.

The molecules used in the training set, for example, may comprise members of target shape groups obtained by the process for obtaining a molecular diversity library, as described above. Training data, e.g., the molecular conformations and/or activities of the training set molecules are subsequently used in a learning model which is refined to generate consistent hypotheses to explain the training data. However, in order to make the learning process more efficient, a bootstrap procedure may be employed. This procedure includes finding stable conformers from the structure data, posing the conformers and selecting initial poses from the poses to form an initial training set. After the training set is formed, the set is used in a learning step to refine a system which is then used to predict the activity of a molecule not in the training set.

To increase the predictive power of the method, the training data preferably includes data on a plurality molecules. As known to those skilled in the art, biologically active molecules can take on different shapes known as conformers or conformations defined by the internal torsion angles of the rotatable bonds in the molecule. An active molecule, for example, may be any molecule that binds a member of a lower or higher ordered target shape group, as described above. In order to increase the computational efficiency in learning, however, it is desirable to choose only the conformations that are best in confirming or refuting the learning model. Thus, addition to molecules that bind members of a lower or higher ordered target shape group, molecules which have weak binding activities may be included in the training set. The model may be designed to account for the relative binding efficiencies of the included molecules by appropriately weighting the importance of each structure within the model.

The first step in this selection involves posing the molecule. A pose of a molecule is defined by its conformation (internal torsion angles of the rotatable bonds) and orientation (three rigid rotations and translations). This mathematically defines the pose or geometrical conformation of the molecule. First, a conformer of an active molecule is chosen and its

pose is first fixed. The initial pose or conformation may be determined from empirical measurements or ab inito calculations. In chemical terms, this is analogous to permitting the molecule to rotate, translate and alter its conformation to achieve its best possible fit to the binding site. The rotation, translation and alteration in the internal torsion angles of the rotatable bonds in a molecule is referred to herein as reposing of the molecule. In other words, since the fixed pose of a molecule known to have high activity is used as the reference for reposing the remaining molecules, this crudely simulates the process of reposing the other molecules to achieve the best possible fit to the binding site. The above-described process can be performed using a number of software packages available commercially, such as, for example, Catalyst from BioCAD, Foster City, Calif., and Batchmin available from Columbia University, New York City, N.Y.

The following example presents a non-limiting example of neural networks applied to conformational data of molecules. As understood by one of ordinary skill in the art, such models could also include empirical activity data such as, for example a relative binding efficiency. Alternatively, or in combination, the model may also include a combination of empirical conformational data and theoretical conformational data, such as, that derived from ab initio molecule structure calculations. The learning process now begins with a selection of only some of the poses to be in the training set. In other words, if a molecule has more than one conformation, poorer matches may be dropped for computational efficiency in the subsequent learning process. In making the selection, various properties of the molecules in the data set known to chemists may be used, including physical and chemical properties such as shape, electrostatic interaction, solvation and biophysical properties, such as, for example, a binding efficiency to a predetermined target.

Before the selected poses may be used for training, the relevant features of these poses are first extracted. The COMFA methodology described in U.S. Pat. No. 5,025,388, for example, employs a three-dimensional lattice structure and extracts the relevant features by calculating the steric and electrostatic interaction energies between a probe atom placed at each of the lattice intersections and the molecule.

Another approach involves creating a surface representation of each of the poses and then obtaining a feature value between at least one sampling point and a point on the surface representation of each of the poses. For example, the van der Waals surface of at least several atoms in a predetermined portion of the molecule, such as, for example the active site, may be found. A curved surface having a number of ridges intersections of adjacent van der Waals surfaces results and is a surface representation of the portion of the molecule.

As known to those skilled in the art, the electron density around each atom can be represented as a Gaussian function of distance from the nucleus of the atom where the peak of such Gaussians would more or less coincide with the van der Waals radius of the atom. A surface representation of the portion of the molecule can then be obtained by summing the Gaussian functions for the atoms. The surface representation arrived at using the van der Waals surfaces of the atom has been found to be adequate and easy to find for most purposes for modeling biological and chemical activity whereas the sum of the Gaussian approach gives a scientifically more rigorous representation of such surface. The details of finding the van der Waals surfaces of atoms and calculations involving a surface such as surface are known to those skilled in the art and will not be explained in detail here. Similarly, the Gaussian distributions for the atoms and method for summing them are also known to those skilled in the art and will not be explained in detail here. Other than van der Waals and Gaussian surface representations, other types of surface representation are possible, such as a Connolly surface. See, M. J. Connolly, J. Appl. Cryst., 16, 548 (1983).

5.5.2a Feature Extraction

The feature values, including steric, electrostatic or other feature values may be extracted by first specifying at least one sampling point and then obtaining a feature value between such sampling point and a point on the surface representation of each of the poses. The point may be outside but near the molecular surface and the feature value is extracted by determining, for example, the minimum distance between such sampling point and the surface representation of the pose. For simplicity, a surface representation of a pose determined in the manner above will be referred to simply as the surface of the pose. An electrostatic feature value may be extracted as the electrostatic interaction between a probe atom placed at such sampling point and the pose. Alternatively, the electrostatic feature value may be the sum of the Coulomb force interactions between the probe atom and atoms of the pose surface. The above described approach will be referred to herein as the point-based feature extraction approach. Preferably, a number of sampling points are chosen surrounding the poses. In other words, the same sampling points are used to extract features from each of the poses in the training set. To arrive at a common set of sampling points, one may select the points by reference to the averaged position of the poses in the training set.

5.5.2b Form of the Model

Once features have been extracted for each initial pose in the initial training set, these features are input to a parameterized mathematical model, such as for example, a neural network or principal component regression to produce an activity prediction. For example, let V (M, P) be the vector of n features extracted to represent molecule M in pose P. Let the kth component of this vector be denoted $V(M, P)_k$.

During training, the optimal values for the model parameters are determined. It will be understood that the scope of this invention includes a wide range of mathematical models, including linear models and nonlinear models. In the preferred embodiment, the model has the form:

$$Activity(V(M, P)) = Sigmoid\left[\sum_{j=1}^{m} \mu_j F_j(v_j, V(M, P), \mu, \sigma)\right]$$

where

m is the number of weights

Sigmoid (x)=1/(1+exp(-x))

exp is the exponential function (whose base e is the base of the natural logarithm)

u_j is a real-valued weight and

$$F_j(V_j, V(M, P), \mu, \sigma) = Sigmoid\left[\sum_{i=1}^{n} V_{j_i} G(V(M, P)_i, \mu_i, \sigma_i)\right]$$

$$G(V(M, P)_i, \mu_i, \sigma_i) = \exp\left[-\frac{(V(M, P)_i - \mu_i)^2}{2\sigma_i^2}\right]$$

μ_i a real-valued location parameter

σ_i is a real-valued width parameter

The parameters of this model are:

u_j (j=1 . . . n)

v_{ji} (j-1 . . . n, i=1 . . . m)

μ_i (i-1 . . . m)

σ_i (i-1 . . . m).

In this embodiment, the function G is a Gaussian-like function that will produce large values when the measured feature $V(M, P)_i$ is near to μ_i and smaller values when the measured feature is distant from μ_i. The value of σ_i controls how rapidly the value of G decreases as $V(M, P)_i$ moves away from μ_i.

Given an initial set of training poses, the training process is initialized by providing starting values for each of the parameters. For example, the values of u_j and v_ji may be set to small random positive values in the range from 0.0 to 0.2; μ_i is initialized to be a small amount (1.0) less than the mean of the values of V (M, P)$_i$ for all molecules and poses in the training data set. The value of σ_i is initially set to a value of 0.25. The value of n, the number of intermediate sigmoids, is initialized to 1. If inadequate predictions are obtained, n can be increased and the model re-trained until a sufficient value of n is found.

The discussion in the preceding paragraphs has focused on steric features, but the same mathematical model applied equally well to electrostatic features. The values of μ_i and σ_i for an electrostatic feature i describe an interval ("Box") of desirable or undesirable values for the feature (depending on the values of v_ji). In fact, the same mathematical model is applicable to other biological activity types including but not limited to binding affinity, agonism, potency, receptor selectivity and tissue selectivity.

5.5.2c Training the Model

The training of the model will now be described. For example, the sampling points may be chosen by reference to an average surface representation obtained by averaging the surface representations of the poses in the training set. Thus, if a surface representation is an averaged surface representation of all the poses, then the sampling points are chosen by reference to such surface. The averaging process to obtain the average representation of a set of poses is known to those skilled in the art.

As explained above, an initial set of poses is selected to form the training set in order to train the model. Then the initial values for the parameters n, μ_i,σ_i, v_ji, and u_j are chosen. The feature values of the poses in the training set are extracted as described above. Then the predicted activity of each of the poses in the training set is calculated using the model and the parameter values set initially by using, for example, the equations above. For each molecule, the pose with the highest predicted activity is chosen as the best pose of the molecule. Then the parameter values set initially for feature i are modified to minimize the differences between the predicted and actual activities of preferably only the best poses of the molecules.

When receptor sites are present in the vicinity of the molecules used for training, it is known that the presence of such sites would influence the orientation and conformations of molecules present so that in actual fact, the molecules would repose under such influence to attempt to conform to the pose with the highest activity. It is of course possible to modify

the parameter values in reference to poses in addition to or other than the best poses; all such variations are within the scope of the invention.

If p_j is the predicted activity of a particular pose j and a_j its actual activity, then an error function for the training set of poses can be formed by, for example, the following equation:

$$Error\ Function = \sum_{j=1}^{m} (P_j - a_j)^2$$

where m is the total number of poses (preferably only the best poses) in the set in reference to which the parameter values are to be modified. A wide variety of computational methods may be applied to minimize the error function with respect to the parameters of the model (e.g., u_j, v_ji, μ_i, σ_i, n). Such methods are known to those skilled in the art and will not be described here. In the preferred embodiment, the gradient of the error function with respect to these parameters (except for n) is computed, and gradient descent methods are applied. Other methods such as conjugate gradient, Newton methods, simulated annealing, and genetic algorithms may also be used and are within the scope of the invention.

After the differences between predicted and actual activities of poses (e.g., best poses) have been minimized, such as by minimizing the above error function, such differences are compared to preset thresholds. If the differences are below the preset threshold or thresholds, one concludes that the process has converged and proceeds. If not, then one returns to calculate the predicted activities of poses in the training set by reference to the modified parameter values and again choose the best pose for each molecule having the highest predicted activity. The parameter values are again modified to minimize differences between predicted and actual activities of best poses. This loop is repeated until the differences are found to be below preset threshold or thresholds and the same best poses are chosen every time.

Then the molecules are reposed to maximize their activities and from the possible poses after the reposing, poses are chosen to form a new training set. Instead of reposing the molecules, it is possible to simply re-select from the initial set of poses to form the training set of poses. However, it is believed to be preferable to repose the molecules in order to form a new training set. The new training set is compared to the prior training set to see whether the changes to the poses are below certain set threshold or thresholds. If the changes are found to be below the threshold(s), then the process of training the model is completed and one proceeds to the prediction step. If the changes to the poses are not below

the threshold or thresholds, then one returns to extract features from the training set as described above.

Since the orientation and conformation of the poses may have changed, these new poses will have different feature values from those in the original training set. Therefore, the feature extraction step needs to be repeated. A molecule may be reposed by first re-orienting the molecule with respect to the sampling points. Then the internal torsion angles of the rotatable bonds are altered to re-conform the molecule to again best fit the surface portions of the molecule

The above-described process makes good use of the salient feature of poses of inactive as well as active molecules. The above-described reposing process with aligned and conformed poses of active molecules to maximize the agreement of the observed a predicted activities and to repose the inactive molecules to be in the best position to refute the model. Thus, in order for the model to pass the above described testing process, it will predict the inactivity of poses of inactive molecules even though these have been realigned and reconfirmed to be in the best position to "fool" the model, while at the same time confirming the activity of the active molecules.

Gradient search methods are also used for reposing the training molecules to maximize their predicted activities as functions of the orientation and conformational parameters.

If the extracted features are differentiable functions of the orientation and conformational parameters and the model (as represented by the equations above) is a differentiable function of the values of the extracted features, the chain rule may be applied to compute the gradient of the predicted activity with respect to the orientation and conformational parameters and apply gradient-based search to find poses that maximize predicted activity. However, other kinds of models and other methods of feature extraction may not satisfy this property, in which case other computational methods (e.g., simulated annealing, linear programming) could be applied to find poses that maximize predicted activity. It is understood that the scope of the invention includes all methods for finding such poses.

Instead of reposing the molecules, it is possible to simply re-select the best poses from the original set of poses formed prior to the selection step. Reposing the molecules rather than re-selecting from existing poses may reduce the error of prediction.

The trained model and the ultimate parameter values may then be used to predict the activity of a new molecule with unknown activity. Thus, again, feature values are extracted from the poses of the molecule and the predicted activities of the poses are calculated to

find the best pose with the highest activity. Thus, the model not only enables the user to predict the activity of the molecule not in the training set but also predict its best poses. Its feature values in comparison with the parameter values would indicate which surface portions have the desirable properties in regard to a chemical function and which surface portions have undesirable properties in regard to such function. In fact, the model may be used to search a database of molecules with unknown activity and predict the activities of their poses. Poses of these molecules may be modified to alter their predicted activities.

5.6 Factor Analysis

Factor analysis provides a technique for expressing the behavior of a system of data in terms of orthogonal vectors, which form a basis set to describe the system. Advantageously, a model is not required to determine the variables which are important in describing the behavior of the system of data. Thus, for example, such an approach would allow one to discriminate from among various shape features those features which are important in determining a desired activity of a molecule. Factor analysis techniques are known in the art and are described, for example, in Lawton and Silvestre, Technometrics, vol. 13, 1971, pp 617-633 Edmund R. Malinowski *Factor Analysis in Chemistry, 2nd Edition* John Wiley & Sons, New York, 1993, which are incorporated in their entireties by reference.

In a factor analysis approach, variables describing members of a targetshape group centered around one or more target compounds may be arranged in a matrix of data. Input variables may include, but are not limited to structural features determined from spectroscopic or other shape determining techniques as described above, the presence of one or more molecular epitopes, the capacity to bind to an antigen, other compound, or surface, the ability to displace another molecule from a binding site, the ability to catalyze one or more reactions, or the ability to modify the catalytic activity of another molecule.

Factor analysis coupled with self-modeling curve resolution allows the identification of significant components responsible for variation in data to be extracted in lieu of a model describing their behavior.

Initially a matrix D is created. The rows of D contain conformation and/or activity data for molecules in the training set. Each column of D, therefore, corresponds to particular conformation or activity data such as, for example, a binding efficiency, an ability to modify an activity of a receptor molecule, a molecular radius, a ability to catalyze a chemical reaction, a bond length, a bond angle, or a relative or absolute distance between predetermined nuclei in each molecule. One of ordinary skill in the art understands that a

variety of conformational and activity data can be used to describe a molecule and each is applicable with the present invention.

To find the correlated variation among the conformation or activity data and to identify the parameters that are best able to describe the activity of the molecules, a covariance matrix Z is formed from the matrix, D:

$$Z = D^T D$$

where Z is a j by j square matrix whose rows describe the correlation between the columns of D. The covariance matrix may be diagonalized by finding a matrix Q such that

$$ZQ = \lambda Q$$

where Q is j by j a matrix of eigenvectors and λ is a diagonal matrix of eigenvalues. The eigenvectors in Q are abstract representations of the variation across the columns of D. The magnitude of the jth eigenvalue, λ_j, indicates the amount of variation in the data described by the jth eigenvector.

If the variation in the data originates only from the n distinguishable components, then the rank of Z (in the absence of noise) would be n, and linear combinations of the n nonzero eigenvectors describe all of the variation across the columns in D. However, due to random experimental error, c - n additional eigenvectors arise from decomposition of Z. These eigenvectors are dominated by experimental noise and their removal does not significantly impact analysis of the correlated behavior of the data. Furthermore, since the number of components, which are needed to describe the behavior of the molecules, in the data is not generally known a priori, determining n can be a significant step toward identifying molecular conformational parameters and activities best suited for identify promising drug leads. Both the relative magnitude of the eigenvalues and the shape of the eigenvectors may be used to estimate of the number of components.

A number of approaches that rely on the relative magnitudes of the eigenvalues have been developed for estimating n when the experimental variance is unknown. The method of reduced eigenvalues, REV, was proposed in 1987 by Malinowski. The jth reduced eigenvalue is given by

$$REV_j = \lambda_j / (r - j + 1)(c - j + 1)$$

The reduced eigenvalue ratio REV_{j+1}/REV_j may be calculated by eq A.3 for c - 1 eigenvalues. According to the theory of this reduced eigenvalue representation, this ratio will be significantly greater than unity only for the n components that describe variation in the data exceeding the noise. The IND function, also developed by Malinowski, should reach a minimum when the correct number of components are included.

$$IND = \left[\sum_{j=n+1}^{c} \frac{\lambda_j}{r\,(c - n)^5} \right]^{1/2}$$

Empirically, the IND function has been found to reach a less convincing minimum than the relative change observed for the eigenvalue ratio, REV.

In general, if the training set includes a large number of molecules, as would be possible with the present invention, the information spanning the row space of D is richer due to the larger number of points along the rows *i.e.* the molecules in the training set and the greater response variation between molecules. A matrix of eigenvectors spanning the row space of D is obtained by projecting the data onto the eigenvectors, q_j

$$U = DQ$$

where U is a r x c matrix whose rows contain the dot product of the jth elgenvector with the ith row in D. Dividing the jth column of U by the square root of the eigenvalue produces a matrix of normalized eigenvectors which are equivalent to the jth eigenvectors obtained by decomposition of the covariance matrix, DD^T.

Following the determination of the number of components, U and Q are truncated to produce an r by n matrix, U and a c by n matrix, Q containing significant eigenvectors in the spectral and intensity-response dimensions. If the number of components was correctly determined, the retained eigenvectors contain all the information required to reproduce the correlated behavior of the data within the limits of the experimental noise of determining each point in D.

Once factor analysis has identified activities or conformational parameters important in describing the desired activity of molecules in the training set, these activities or conformational parameters can be used to design a new molecule for synthesis or to scan existing data bases for molecules having similar activities or conformational parameters that may prove to be suitable drug leads. As with any of the methods used in the present invention, once a putative structure or molecule is identified one or more variants may be

generated to obtain a focused diversity library of molecules having a structure at least somewhat similar to the putative structure. In this fashion, the present invention, which combines rational drug design an molecular diversity allows "pseudo-random" variants to be obtained, which are focused in a desirable region of molecular shape space. Moreover, the target shape groups of the present invention allow the inclusion of molecules in the modeling step which bind so weakly to the target molecule as to remain undetected but which nonetheless provide relevant information to the model increasing the predictability.

5.7 Principal Component Analysis

As understood in the art, principal component analysis (PCA) provides a robust approach to modeling empirical data, which can be utilized with the present invention. Principal component analysis is similar to factor analysis in requiring a matrix representation of the data but differs in that it also involves a regression step. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, NewYork, 1986. Richard A. Reyment, et al *Applied Factor Analysis in the Natural Sciences*, Cambridge Univ Press, New York, 1996, describe PCA approaches to data modeling and are included in their entireties by reference.

5.8 Computer Systems

FIG. 2 discloses a representative computer system 810 in conjunction with which the embodiments of the present invention may be implemented. Computer system 810 may be a personal computer, workstation, or a larger system such as a minicomputer. However, one skilled in the art of computer systems will understand that the present invention is not limited to a particular class or model of computer.

As shown in FIG. 3, representative computer system 810 includes a central processing unit (CPU) 812, a memory unit 814, one or more storage devices 816, an input device 818, an output device 820, and communication interface 822. A system bus 824 is provided for communications between these elements. Computer system 810 may additionally function through use of an operating system such as Windows, DOS, or UNIX. However, one skilled in the art of computer systems will understand that the present invention is not limited to a particular operating system.

Storage devices 816 may illustratively include one or more floppy or hard disk drives, CD-ROMs, DVDs, or tapes. Input device 818 comprises a keyboard, mouse, microphone, or other similar device. Output device 820 is a computer monitor or any other known computer

output device. Communication interface 822 may be a modem, a network interface, or other connection to external electronic devices, such as a serial or parallel port.

Exemplary configurations of the representative computer system 810 include client-server architectures, parallel computing, distributed computing, the Internet, etc. However, one skilled in the art of computer systems will understand that the present invention is not limited to a particular configuration.

The present invention is not to be limited in scope by the specific embodiments described herein, which are intended as single illustrations of individual aspects of the invention, and functionally equivalent methods and components are within the scope of the invention. Indeed, various modifications of the invention, in addition to those shown and described herein will become apparent to those skilled in the art from the foregoing description and accompanying drawings. Such modifications are intended to fall within the scope of the appended claims. Various references including patent applications, patents, and other publications, are cited herein, the disclosures of which are incorporated by reference in their entireties.

Claims

What is claimed is:

1. A method for predicting a property of a molecule comprising the steps of:

obtaining an initial odd set of molecules that bind at least one molecule belonging to an origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

selecting a training set comprising a subset of the even set of molecules;

determining a conformation for each of the training set molecules;

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subset of the even set of molecules and wherein the new molecule has a new conformation and the model includes the conformation of at least some of the training set molecules; and

predicting the predetermined property of the new molecule.

2. The method according to claim 1 comprising the further steps of:

selecting a training set comprising a subset from the odd set of molecules; and

repeating the determining, constructing, and predicting steps wherein the model further comprises the conformation for each molecule in the odd subset.

3. The method of claim 1 wherein the constructing a model step further comprises the steps of:

predicting a predetermined property of at least one molecule assigned to one of the subsets;

conditionally modifying the model in response to a difference between said predicted predetermined property and an empirical estimate of said predicted property; and

repeating said predicting and conditionally modifying steps until said difference reaches a predetermined value.

4. The method of claim 1 wherein the predetermined property is the ability to bind to at least one predetermined molecule and the empirical estimate is determined from a binding assay.

5. The method of claim 1 wherein the model comprises at least one of a neural network, a factor analysis, or a principal components analysis.

6. The method of claim 5 wherein the model is a neural network comprising:

a plurality of layers each having at least one node wherein the plurality of layers include a first layer having at least one node coupled to an input value and a second layer having at least one node coupled to a plurality of nodes of said first layer; and

the first layer having at least one node with a first transfer function and the second layer having at least one node with a second transfer function.

7. The method of claim 5 wherein the conformation is determined by at least one of x-ray crystallography, nuclear magnetic resonance, or molecular modeling.

8. The method of claim 1 wherein the conformation comprises at least one of an absolute positions of atomic nuclei in each molecule, a relative position of atomic nuclei in each molecule, an electron density distribution, a bond angle, a bond length, or a van der Waals radii of atoms in the molecule.

9. The method of claim 1 further comprising the step of searching a conformational data base for molecules having a conformation similar to the new conformation.

10. The method of claim 1 further comprising the step of synthesizing at least a portion of the new molecule.

11. The method of claim 9 wherein the new molecule comprises at least one of DNA, RNA, a peptide, a polypeptide, or a small molecule.

12. The method of claim 10 further comprising the steps of:

producing one or more first variants of the new molecule that are at least somewhat similar to the new molecule; and

selecting one or more of the first variants having at least one desired characteristic.

13. The method according to claim 12 wherein the first variants comprise a stochastic sequence of polynucleotides.

14. The method according to claim 10 further comprising raising antibodies against the new molecule.

15. A method for predicting a property of a molecule comprising the steps of:

obtaining an initial odd set of molecules that bind at least one molecule belonging to an origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

obtaining an odd set of molecules that bind at least one molecule belonging to said even set of molecules;

repeating said obtaining an odd set of molecules and said obtaining an even set of molecules steps to generate a sequence of odd and even sets of molecules wherein the molecules in each of said sets bind to at least one of the molecules in a preceding one of the sets in the sequence; and

selecting a training set comprising an even subset from each of at least two even sets of molecules;

determining a conformation for each molecule in each of said subsets;

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subsets of molecules wherein the model comprises the conformation of at least some of the molecules from each even subset; and

predicting a predetermined property of the new molecule.

16. The method of claim 15 further comprising the steps of:

selecting a training set comprising a subset from each of at least two odd sets of molecules; and

repeating the determining, constructing, and predicting steps wherein the model further comprises the conformation for each molecule in each of odd subsets.

17. The method of claim 16 wherein the constructing a model step further comprises the steps of:

predicting a predetermined property of at least one molecule assigned to one of the subsets;

conditionally modifying the model in response to a difference between said predicted predetermined property and an empirical estimate of said predicted property; and

repeating said predicting and conditionally modifying steps until said difference reaches a predetermined value.

18. The method of claim 17 wherein the predetermined property is the ability to bind to at least one predetermined molecule.

19. The method of claim 18 wherein the model comprises at least one of a neural network, a factor analysis model, a principal components analysis model, or an independent component analysis model.

20. A method for predicting a property of a molecule comprising the steps of:

selecting a first origin set of molecules;

obtaining an initial odd set of molecules that binds at least one molecule belonging to the first origin set of molecules;

obtaining an even set of molecules that bind at least one molecule belonging to said odd set of molecules;

obtaining an odd set of molecules that bind at least one molecule belonging to said even set of molecules;

repeating said obtaining an odd set of molecules and said obtaining an even set of molecules steps to generate a sequence of odd and even sets of molecules wherein the molecules in each of said sets bind to at least one of the molecules in a preceding one of the sets in the sequence;

selecting a second origin set of molecules and repeating said obtaining an initial odd set, obtaining an even set, obtaining an odd set and said repeating steps to generate a second sequence of odd and even sets of molecules;

selecting a training set comprising an even subset from each of at least two even sets of molecules belonging to the first and second sequences;

determining a conformation for each molecule in each of said subsets;

constructing a model for predicting a predetermined property of at least one new molecule not assigned to the subsets of molecules wherein the model comprises the conformation of at least some of the molecules from each even subset; and

predicting a predetermined property of the new molecule.

21. The method of claim 20 wherein the predetermined property comprises the ability of the new molecule to bind to each of at least two predetermined molecules.

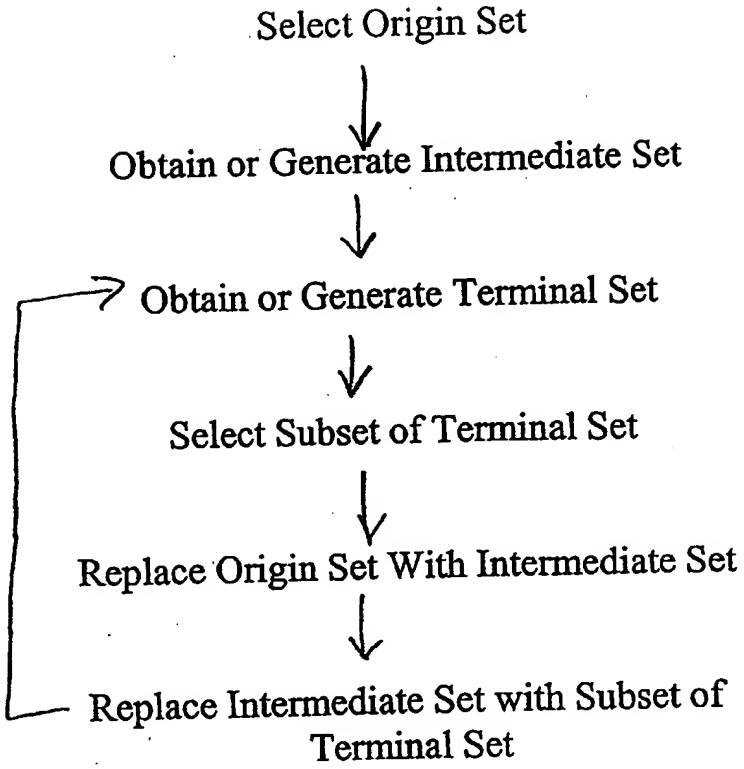
Select Origin Set
Obtain or Generate Intermediate Set
Obtain or Generate Terminal Set
Select Subset of Terminal Set
Replace Origin Set With Intermediate Set
Replace Intermediate Set with Subset of Terminal Set

Figure 1

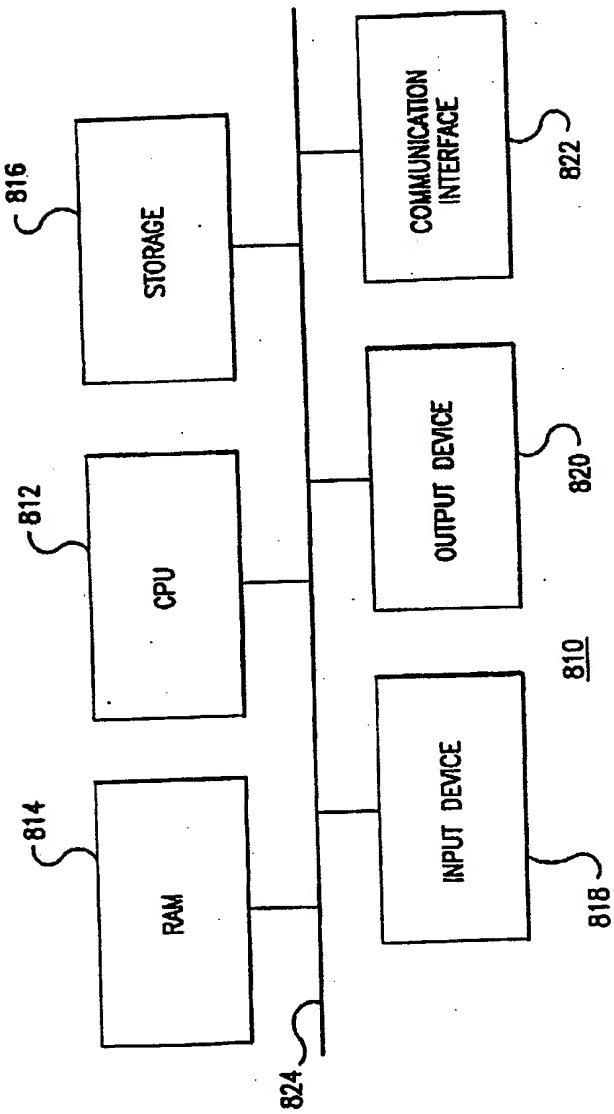
816
STORAGE
812
CPU
814
RAM
824
COMMUNICATION INTERFACE
822
OUTPUT DEVICE
820
810
INPUT DEVICE
818

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/10484

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :C12Q 1/00, 1/68; C12P 19/24, 21/06; C07H 21/02; G01N 33/48, 33/50
US CL :435/4, 6, 94, 68.1, 91.1; 536/25.3; 702/19, 20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/4, 6, 94, 68.1, 91.1; 536/25.3; 702/19, 20

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

MEDLINE, BIOSIS, SCISEARCH, CAPLUS, WEST
molecule, target, molecular diversity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,837,458 A (MINSHULL et al) 17 November 1998, see entire document.	1-21
A	CAPORALE L.H. Chemical ecology: A view from the pharmaceutical industry. Proc. Natl. Acad. Sci., USA. January 1995, Vol. 92, pages 75-82, see entire document.	1-21

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 14 JUNE 2000	Date of mailing of the international search report 02 AUG 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer STEPHEN SIU Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)*